Contrastive Learning for Efficient Reranking in Task 14B of the CLEF 2026 BioASQ Track

University of Amsterdam at the CLEF 2026 BioASQ Track

Jan Bakker[1], Jaap Kamps[1]

[1]*Institute for Logic, Language and Computation (ILLC), University of Amsterdam, The Netherlands*

Abstract

This paper describes the University of Amsterdam's participation in BioASQ Task 14B Phase A on the Large-Scale Semantic Indexing of Biomedical Abstracts. We took a general DeBERTa-based reranker and fine-tuned it on the BioASQ train set via contrastive learning, using abstracts that were marked relevant to a given question as positives, and unmarked abstracts that were retrieved by SPLADE-v3 as negatives. Although our straightforward approach could be expected to improve the reranker's ability to handle biomedical queries and abstracts, the preliminary evaluation results indicate little to no improvement in retrieval effectiveness. We conclude that other strategies might yield better results.

Keywords

Biomedical Semantic Indexing, Efficient Reranking, Contrastive Learning

1. Introduction

Suppose that you are a biomedical scientist who has a specific biomedical question. For example, you would like to know the answer to the following question: *which cardiovascular events are associated with ALOX12 gene polymorphisms*? In that case, having an efficient way to retrieve the abstracts of biomedical articles that are relevant to this question would be of great utility. You could read the retrieved abstracts and corresponding articles to formulate an answer to your question. Additionally, you could use a Large Language Model (LLM) to generate an answer based on the retrieved abstracts [1]. The latter technique is known as Retrieval-Augmented Generation (RAG).

The aim of the BioASQ Challenge is to push the research frontier towards systems that respond directly to the information needs of biomedical scientists [2]. In Task 14B, participants receive batches of biomedical questions in English. They have to respond with retrieved articles (Phase A) and generated answers (Phase A+/B), which are then matched against a set of initial references created by experts. Later, these experts will manually assess the responses of all teams and update the references accordingly.

This paper describes the University of Amsterdam's participation in BioASQ Task 14B Phase A on Large-Scale Biomedical Semantic Indexing. In summary, we took a general efficient DeBERTa-based reranker and fine-tuned it on the BioASQ train set via contrastive learning. We describe our approach in more detail in Section 2. Subsequently, we present the preliminary results in Section 3. We end with a discussion of these results and our conclusions in Section 4. Our fine-tuning code is available at https://github.com/JanB100/bioasq14b.

2. Experimental setup

Following the guidelines for BioASQ Task 14B Phase A, we retrieved relevant articles from the PubMed annual baseline for 2026 based on their abstracts. Upon downloading, we reconstructed each abstract from its XML record and prepended the title of the corresponding article. The resulting dataset comprised 28M PubMed abstracts and their identifiers.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

j.bakker@uva.nl (J. Bakker); kamps@uva.nl (J. Kamps)

0009-0002-9085-8491 (J. Bakker); 0000-0002-6614-0087 (J. Kamps)

We submitted four runs per batch. For our first run, we only used SPLADE-v3 as retriever [3]. SPLADE is a sparse bi-encoder model that maps queries and passages alike to a 30522-dimensional sparse vector space, and efficiently calculates their similarity using the dot product [4]. It was trained on the large-scale MS MARCO passage ranking dataset, which contains many web search queries and passages with human annotated relevance labels [5]. SPLADE encodes inputs of up to 512 tokens, thereby truncating approximately 8% of all abstracts. We retrieved 1000 abstracts per query.

For our second run, we additionally used the DeBERTa-based reranker[1] that Lassance and Clinchant [6] trained for their participation in the TREC 2022 Deep Learning challenge [7]. It is a cross-encoder model that processes each query-passage pair together and outputs a ranking score, and it is relatively efficient compared to other reranking models due to its small size. Lassance and Clinchant [6] trained their reranker on the MS MARCO passage ranking dataset through contrastive learning, using negatives from SPLADE-v3. That is, inspired by Gao et al. [8], they trained their reranker to effectively discriminate between truly relevant passages and the top irrelevant passages retrieved by SPLADE-v3. We used the resulting reranker to rerank the top 100 retrieved abstracts per query, truncated again to 512 tokens.

Furthermore, we fine-tuned this reranker on the BioASQ train set [9] using the codebase[2] of Gao et al. [8]. In doing so, we took some inspiration from the approach of Almeida et al. [10] to BioASQ Task 12B Phase A. We used abstracts that were marked relevant to a given question as positives, and random unmarked abstracts that were retrieved by SPLADE-v3 as negatives. Since the relevance labels in the BioASQ train set are based on specific versions of the PubMed annual baseline, we only used abstracts that were published in the years before a given question was introduced in BioASQ as negatives.

For the third run, we fine-tuned the reranker on all training instances created before 2025. For the fourth run, we instead fine-tuned the reranker on a recent subset of training instances created from 2020 to 2024 (BioASQ 8-12). We then used these models to rerank the top-100 abstracts per query retrieved by SPLADE-v3. We fine-tuned our rerankers on 4 GPU A100s in half-precision. We set the learning rate to 1e-5 and the batch size per device to 8, with 1 positive and 7 negatives per batch. We set the number of epochs to 2 for the third run and 4 for the fourth. Thus, fine-tuning took about 15 minutes per model.

We used the BERGEN library [11] to perform inference.

3. Results

Table 1 shows the mean retrieval scores of our systems on the combined test sets for BioASQ 13B, based on the top-10 abstracts retrieved per question. In addition, Table 2 shows the preliminary retrieval scores of our systems on test batches 2, 3, and 4 of BioASQ 14B. The evaluation metrics are mean precision, mean recall, mean F-measure, mean average precision, and geometric mean average precision.

The references for the test sets of BioASQ 13B were updated by experts based on the runs submitted last year. This is why the scores in Table 1 are higher and more representative than the preliminary scores in Table 2. However, if our systems identified any relevant articles that the systems of other participants missed, then their actual performance would be even higher. Therefore, the manual assessment of the runs submitted to BioASQ 14B Phase A will paint a more complete picture.

	System	Precision	Recall	F-measure	MAP	GMAP
Batch 2	Splade-v3	30.11	22.24	16.90	22.31	0.03
	Debertav3	37.55	26.38	20.95	29.92	0.07
	Fully fine-tuned	34.32	25.19	19.45	26.06	0.06
	Recently fine-tuned	37.88	26.66	21.29	30.10	0.07

Table 1
Mean document retrieval scores on the combined test sets of BioASQ 13B. The top-10 retrieved abstracts per question are evaluated.

[1] https://huggingface.co/naver/trecdl22-crossencoder-debertav3
[2] https://github.com/luyug/Reranker

	System	Precision	Recall	F-measure	MAP	GMAP
Batch 2	Splade-v3	3.75	13.68	5.60	9.28	0.02
	Debertav3	5.00	18.81	7.43	12.41	0.04
	Fully fine-tuned	4.25	16.70	6.37	12.18	0.02
	Recently fine-tuned	5.25	19.51	7.80	10.92	0.04
Batch 3	Splade-v3	3.67	16.13	5.75	9.26	0.02
	Debertav3	5.33	21.90	8.14	13.95	0.05
	Fully fine-tuned	4.67	19.35	7.10	12.26	0.03
	Recently fine-tuned	5.17	23.27	8.04	14.11	0.06
Batch 4	Splade-v3	4.33	18.52	6.36	10.35	0.03
	Debertav3	6.00	25.44	8.96	15.03	0.07
	Fully fine-tuned	4.83	20.67	7.35	14.47	0.04
	Recently fine-tuned	5.83	25.10	8.73	14.20	0.06

Table 2
Preliminary mean document retrieval scores on test batches 2, 3, and 4 of BioASQ 14B. The top-10 retrieved abstracts per question are evaluated.

For now, we observe that our fine-tuned rerankers did not obtain much higher retrieval scores than the baseline reranker. Although the reranker that was fine-tuned on a recent subset of BioASQ achieved slightly higher scores in some evaluations, we find that this effect is not substantial. In contrast, we also observe that reranking the top abstracts retrieved by SPLADE-v3 with a DeBERTa-based cross-encoder leads to clear improvements in retrieval effectiveness. Compared to the top systems submitted to BioASQ 13B Phase A, our scores are relatively low.[3] However, the scores obtained by these systems are a result of manual assessment and could be based on less efficient retrieval strategies.

4. Conclusion

In this paper, we described the University of Amsterdam's participation in BioASQ Task 14B Phase A on the Large-Scale Semantic Indexing of Biomedical Abstracts. In summary, we took a general efficient DeBERTa-based reranker and fine-tuned it on the BioASQ train set via contrastive learning. Although our straightforward approach could be expected to improve the reranker's ability to handle biomedical queries and abstracts, the preliminary evaluation results indicate little to no improvement in retrieval effectiveness. We conclude that other strategies might yield better results.

Acknowledgments

Jan Bakker and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull in order to: Grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

[3] https://participants-area.bioasq.org/results/13b/phaseA/

References

[1] J. Bakker, J. Kamps, Generating Long-Form Answers to Biomedical Questions: Effectiveness and Efficiency, in: TREC 2025 Proceedings, 2026. URL: https://trec.nist.gov/pubs/trec34/papers/UAmsterdam.biogen.pdf.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[3] C. Lassance, H. Déjean, T. Formal, S. Clinchant, SPLADE-v3: New baselines for SPLADE, 2024. URL: https://arxiv.org/abs/2403.06789. arXiv:2403.06789.

[4] T. Formal, B. Piwowarski, S. Clinchant, Splade: Sparse lexical and expansion model for first stage ranking, 2021. URL: https://arxiv.org/abs/2107.05720. arXiv:2107.05720.

[5] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco: A human generated machine reading comprehension dataset, 2018. URL: https://arxiv.org/abs/1611.09268. arXiv:1611.09268.

[6] C. Lassance, S. Clinchant, Naver Labs Europe (SPLADE) @ TREC Deep Learning 2022, 2023. URL: https://arxiv.org/abs/2302.12574. arXiv:2302.12574.

[7] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, J. Lin, E. M. Voorhees, I. Soboroff, Overview of the trec 2022 deep learning track, 2025. URL: https://arxiv.org/abs/2507.10865. arXiv:2507.10865.

[8] L. Gao, Z. Dai, J. Callan, Rethink training of bert rerankers in multi-stage retrieval pipeline, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2021, p. 280–286. URL: https://doi.org/10.1007/978-3-030-72240-1_26. doi:10.1007/978-3-030-72240-1_26.

[9] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for biomedical question answering, Scientific Data 10 (2023) 170. URL: https://doi.org/10.1038/s41597-023-02068-4. doi:10.1038/s41597-023-02068-4.

[10] T. Almeida, R. A. A. Jonker, J. A. Reis, J. R. Almeida, S. Matos, https://ceur-ws.org/vol-3740/paper-05.pdf, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271772518.

[11] D. Rau, H. Déjean, N. Chirkova, T. Formal, S. Wang, S. Clinchant, V. Nikoulina, BERGEN: A benchmarking library for retrieval-augmented generation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 7640–7663. URL: https://aclanthology.org/2024.findings-emnlp.449/. doi:10.18653/v1/2024.findings-emnlp.449.